

On the ℓ_1 - ℓ_q Regularized Regression

Han Liu^{1,2}, Jian Zhang³

¹Machine Learning Department, ²Statistics Department
Carnegie Mellon University, Pittsburgh, PA, 15213

³Department of Statistics
Purdue University, West Lafayette, IN, 47907

February 11, 2008

ABSTRACT

In this paper we consider the problem of grouped variable selection in high-dimensional regression using ℓ_1 - ℓ_q regularization ($1 \leq q \leq \infty$), which can be viewed as a natural generalization of the ℓ_1 - ℓ_2 regularization (the group Lasso). The key condition is that the dimensionality p_n can increase much faster than the sample size n , i.e. $p_n \gg n$ (in our case p_n is the number of groups), but the number of relevant groups is small. The main conclusion is that many good properties from ℓ_1 -regularization (Lasso) naturally carry on to the ℓ_1 - ℓ_q cases ($1 \leq q \leq \infty$), even if the number of variables within each group also increases with the sample size. With fixed design, we show that the whole family of estimators are both estimation consistent and variable selection consistent under different conditions. We also show the persistency result with random design under a much weaker condition. These results provide a unified treatment for the whole family of estimators ranging from $q = 1$ (Lasso) to $q = \infty$ (iCAP), with $q = 2$ (group Lasso) as a special case. When there is no group structure available, all the analysis reduces to the current results of the Lasso estimator ($q = 1$).

Keywords: ℓ_1 - ℓ_q regularization, ℓ_1 -consistency, variable selection consistency, sparsity oracle inequalities, rates of convergence, Lasso, iCAP, group Lasso, simultaneous Lasso

I. INTRODUCTION

We consider the problem of recovering a high-dimensional vector $\beta^* \in \mathbb{R}^{m_n}$ using a sample of independent pairs $(X_{1\bullet}, Y_1), \dots, (X_{n\bullet}, Y_n)$ from a multiple linear regression model, $Y = X\beta^* + \epsilon$. Here Y is the $n \times 1$ response vector and X represents the observed $n \times m_n$ design matrix whose i -th row vector is denoted by $X_{i\bullet}$. β^* is the true unknown coefficient vector that we want to recover, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ is an $n \times 1$ vector of i.i.d. noise with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In this paper we are interested in the situation where all the variables are naturally partitioned into p_n groups. Grouped variables often appear in real world applications. For example, in many data mining problems we encode categorical variables using a set of dummy variables and as a result they form a group. Another example is additive model, where each component function can be represented using its basis expansions which can be treated as a group. Suppose the number of variables in the j -th group is represented by d_j , then by definition we have $m_n = \sum_{j=1}^{p_n} d_j$. We can rewrite the above linear model as

$$Y = X\beta^* + \epsilon = \sum_{j=1}^{p_n} X_j\beta_j^* + \epsilon \quad (1.1)$$

where X_j is an $n \times d_j$ matrix corresponding to the j -th group (which could be either categorical or continuous) and β_j^* is the corresponding $d_j \times 1$ coefficient subvector. Therefore, we have $X = (X_1, \dots, X_{p_n})$ and $\beta^* = (\beta_1^{*T}, \dots, \beta_{p_n}^{*T})^T$. All predictors and the response variable are assumed to be centered at zero to simplify notation. Furthermore, we use $X_{\underline{j}}$ to represent the j -th column in the design matrix X and assume that all columns in the design matrix are standardized, i.e. $\frac{1}{n}\|X_{\underline{j}}\|_{\ell_2}^2 = 1, \underline{j} = 1, \dots, m_n$. Similar to the notation of $X_{\underline{j}}$, we denote $\beta_{\underline{j}}^*$ ($\underline{j} = 1, \dots, m_n$) to be the j -th individual element of the vector β^* . Since we are mainly interested in the high-dimensional setting, we allow the number of groups p_n to increase as the number of examples n increases and our results mainly focus on the case where $p_n \gg n$. Furthermore, we also allow the group size d_j to increase with n at a rate $d_j = o(n)$ and define $\bar{d}_n = \max_j d_j$ to be the upper bound of the group size for a fixed n . In the rest of the paper we will suppress the subscript n when there is no confusion.

In order to obtain a reliable estimation of β^* when $p_n \gg n$, the key assumption is that the true coefficient vector β^* is *sparse*. Denote $S = \{j : \|\beta_j^*\|_{\ell_\infty} \neq 0, j = 1, \dots, p_n\}$ to be the set of group indices and let $s_n = |S|$ to be the cardinality of the set S , we also denote β_S^* to be the vector concatenating all subvectors β_j^* 's for $j \in S$. The sparsity assumption means that $s_n \ll p_n$. Therefore, even if β^* has a very high dimension, the only effective part is β_S^* while the remaining part $\beta_{S^c}^* = \mathbf{0}$. Our task is to select and recover the nonzero groups of variables corresponding to the index set S .

Sparsity has a long history of successes in solving such high-dimensional problems. Without considering the group structure, there exist many classical methods for variable selection, such as AIC (Akaike, 1973), BIC (Schwarz, 1978), Mallows's C_p (Mallows, 1973), etc. Although these methods have been proven to be theoretically sound and have been shown to

perform well in practice, they are only computationally feasible when the number of variables is small. Recently, more attention has been focused on the ℓ_1 -regularized least squares (Lasso) estimator (Tibshirani, 1996; Chen et al., 1998) which is defined as

$$\widehat{\beta}^{\lambda_n} = \arg \min_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_{\ell_2}^2 + \lambda_n \|\beta\|_{\ell_1} \right\} \quad (1.2)$$

where λ_n is the regularization parameter for the ℓ_1 -norm of the coefficients β , while $\widehat{\beta}^{\lambda_n}$ means the Lasso solution when λ_n is used for regularization. In the following, we will suppress the superscript if not confusion is caused. Lasso can be formulated as a quadratic programming problem and the solution can be solved efficiently (Osborne et al., 2000; Efron et al., 2004). Its asymptotic properties for fixed dimensionality have been studied in (Fu and Knight, 2000). For high dimensional setting, Greenshtein and Ritov (2004) prove that Lasso estimator is persistent, in the sense that, when constrained in a class, the predictive risk of the Lasso estimator converges to the risk obtained by the oracle estimator in probability. However, recent studies (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2007; Zou, 2006) show that the Lasso estimator is not in general variable selection consistent, which means that in general the correct sparse subset of the relevant variables can not be identified even asymptotically. In particular, in (Zhao and Yu, 2007; Wainwright et al., 2006), it is shown that in order for Lasso to be variable selection consistent, the so-called irrepresentable condition has to be satisfied. Zou (2006) propose the adaptive Lasso and show that by using adaptive weights for different variables, the ℓ_1 penalty can lead to variable selection consistency. In terms of estimation, it has been show in Meinshausen and Yu (2006) that under weaker conditions, the Lasso estimator is ℓ_2 -consistent for high-dimensional setting where the total number of variables can grow almost as fast as $\exp(n)$. Under a stronger assumption, Bunea et al. (2007a) further proves the sparsity oracle inequalities for the Lasso estimator using fixed design, which bounds the ℓ_2 -norm of the predictive error in terms of the number of non-zero components of the oracle vector. Such results can be used applied to nonparametric adaptive regression estimation and to the problem of aggregation of arbitrary estimators. Parallel to the fixed design result, a similar result for the random design can be found in (Bunea et al., 2007b). A more recent result from (Bickel et al., 2007) refine similar oracle inequalities using weaker assumptions. All these results show that for sparse linear models, Lasso can overcome the curse of dimensionality even when facing increasing dimensions.

When variables are naturally grouped together, it is more meaningful to select variables at a group level instead of individual variables, as can be seen from previous examples. A general strategy for grouped variable selection is to use block ℓ_1 -norm regularization. For variables within each block (group), an ℓ_q norm is applied, and different blocks are then combined by an ℓ_1 norm (therefore the name ℓ_1 - ℓ_q regularization). One such example is the group Lasso (Yuan and Lin, 2006), which is an extension of Lasso for grouped variable and can be viewed an ℓ_1 - ℓ_2 regularized regression. Other works related to grouped variable selection include the iCAP estimator (Zhao et al., 2008), which can be viewed as an ℓ_1 - ℓ_∞ regularized regression, and group logistic regression (Meier et al., 2007), etc. Using random design,

Meier et al. (2007) proved the estimation consistency result for group Lasso with Lipschitz type loss functions. Also with random design, Bach (2007) derived a similar irrepresentable condition as in (Zhao and Yu, 2007) and proved the variable selection consistency result for group Lasso. However, to the best of our knowledge, there isn't corresponding result for estimation and variable selection consistency for the group Lasso and iCAP estimators using fixed design, nor the persistency results using random design. There is also no systematic theoretical treatment for the whole family of the more general ℓ_1 - ℓ_q regularized regression with $1 \leq q \leq \infty$.

Our work tries to bridge this gap and provide a unified treatment of ℓ_1 - ℓ_q regularized regression for the whole range from $q = 1$ to $q = \infty$. The main conclusion of our study is that many good properties from ℓ_1 -regularization (Lasso) naturally carry on to the ℓ_1 - ℓ_q cases ($1 \leq q \leq \infty$), even if the number of variables within each group can increase with the sample size n . Using fixed design, when different conditions are assumed, we show that ℓ_1 - ℓ_q estimator is both estimation consistent and variable selection consistent, and if the linear model assumption does not hold, sparsity oracle inequalities for the prediction error could still be obtained under a weaker condition. Using random design, we show that a constrained form of the ℓ_1 - ℓ_q regression estimator is persistent. Our results provide simultaneous analysis to both the iCAP ($q = \infty$) and the group Lasso estimators ($q = 2$). When there is no group structure, all the analysis naturally reduces to the current results of the Lasso estimator ($q = 1$). One interesting application of these results is to analyze the simultaneous Lasso estimator (Turlach et al., 2005), which can be viewed as an ℓ_1 - ℓ_∞ regularized regression using block designs.

The rest of the paper is organized as follows. In Section 2 we first introduce some preliminaries of the ℓ_1 - ℓ_q regularized regression and then describe some characteristics of its solution. In Section 3, we study the variable selection consistency result. In Section 4, we study the estimation consistency and the sparsity oracle inequalities. In Section 5, we study the persistency property. We conclude with some discussion in Section 6.

II. ℓ_1 - ℓ_q REGULARIZED REGRESSION

Given the design matrix X and the response vector Y , the ℓ_1 - ℓ_q regularized regression estimator is defined as the solution of the following convex optimization problem:

$$\hat{\beta}^{\lambda_n} = \arg \min_{\beta} \frac{1}{2n} \|Y - X\beta\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q} \quad (2.1)$$

where λ_n is a positive number which penalizes complex model and q' is the conjugate exponent of q , which satisfies $\frac{1}{q'} + \frac{1}{q} = 1$ (assuming $\frac{1}{\infty} = 0$). The terms $(d_j)^{1/q'}$ are used to adjust the effect of different group sizes. It is easy to see that when $q = 1$, this reduces to the stan-

dard Lasso estimator; when $q = 2$, this reduces to the group Lasso estimator (Yuan and Lin, 2006); when $q = \infty$, this reduces to the ℓ_1 - ℓ_∞ regularized regression estimator, or the iCAP estimator defined in (Zhao et al., 2008).

To characterize the solution to this problem, the following result can be straightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality condition for convex optimization.

Proposition 2.1. *(KKT conditions) A vector $\hat{\beta} = (\hat{\beta}_1^T, \dots, \hat{\beta}_p^T)^T \in \mathbb{R}^{m_n}$, $m_n = \sum_{j=1}^{p_n} d_j$, is an optimum of the objective function in (2.1) if and only if there exists a sequence of subgradients $\hat{g}_j \in \partial \|\hat{\beta}_j\|_{\ell_q}$, such that*

$$\frac{1}{n} X_j^T (X\hat{\beta} - Y) + \lambda_n (d_j)^{1/q'} \hat{g}_j = \mathbf{0}. \quad (2.2)$$

The subdifferentials $\partial \|\hat{\beta}_j\|_{\ell_q}$ is the set of vectors $\hat{g}_j \in \mathbb{R}^{d_j}$ satisfying

If $1 < q < \infty$, then

$$\hat{g}_j = \partial \|\hat{\beta}_j\|_{\ell_q} = \begin{cases} B^{q'}(1) & \text{if } \hat{\beta}_j = \mathbf{0} \\ \left\{ \left(\frac{|\hat{\beta}_{j\ell}|^{q-1} \text{sign}(\hat{\beta}_{j\ell})}{\|\hat{\beta}_j\|_{\ell_q}^{q-1}} \right)_{\ell=1}^{d_j} \right\} & \text{o.w.} \end{cases} \quad (2.3)$$

where $B^{q'}(1)$ denotes the ball of radius 1 in the dual norm, i.e. $1/q + 1/q' = 1$. It's easy to see that $\|\hat{g}_j\|_{\ell_{q'}} \leq 1$ for any j .

If $q = \infty$ then

$$\hat{g}_j = \partial \|\hat{\beta}_j\|_{\ell_\infty} = \begin{cases} B^1(1) & \text{if } \hat{\beta}_j = \mathbf{0} \\ \text{conv}\{\text{sign}(\hat{\beta}_{j\ell}) e_\ell : |\hat{\beta}_{j\ell}| = \|\hat{\beta}_j\|_{\ell_\infty}\} & \text{o.w.} \end{cases} \quad (2.4)$$

where $\text{conv}(A)$ denotes the convex hull of a set A and e_ℓ the ℓ -th canonical unit vector in \mathbb{R}^{d_j} . It's also easy to see that $\|\hat{g}_j\|_{\ell_{q'}} = \|\hat{g}_j\|_{\ell_1} \leq 1$ for all j when $q = \infty$.

If $q = 1$ then

$$\hat{g}_j = \partial \|\hat{\beta}_j\|_{\ell_1} = \{\xi \in \mathbb{R}^{d_j} : \xi_\ell \in \partial |\cdot|(x_\ell), \ell = 1, \dots, d_j\}. \quad (2.5)$$

From proposition 2.1, the ℓ_1 - ℓ_q regularized regression estimator can be efficiently solved even with large n and p_n . For example, blockwise coordinate descent algorithms as in (Zhao et al., 2008) can be easily applied. When $q = 1$ and $q = \infty$, due to fact that feasible parameters are constrained to lie within a polyhedral region with parallel level curves, efficient path algorithm can be developed (Efron et al., 2004; Zhao et al., 2008). At each iteration of the blockwise coordinate descent algorithm, β_j for $j = 1, \dots, p_n$ is updated, with the rest of the coefficients fixed. Coupled with a threshold operator, these algorithms general converge very

fast and exact solution can be obtained. Standard optimization methods, such as interior-point methods (Boyd and Vandenberghe, 2004), can also be directly applied to solve the ℓ_1 - ℓ_q regularized regression problems.

It is well-known (Osborne et al., 2000) that under some conditions, the Lasso can at most select n nonzero variables even in the case $p_n \gg n$. A similar but weaker result can be obtained for the ℓ_1 - ℓ_q regularized regression.

Proposition 2.2. *For the ℓ_1 - ℓ_q regularized regression problem defined in equation (2.1) with $\lambda_n > 0$, there exists a solution $\hat{\beta}^\lambda$ such that the number of nonzero groups $|S(\hat{\beta})|$ is upper bounded by n , the number of given data points, where $S(\hat{\beta}) = \{j : \hat{\beta}_j \neq \mathbf{0}\}$*

Remark 2.3. Notice that the solution to ℓ_1 - ℓ_q regularized regression problem may not be unique especially when $p_n \gg n$ (similar to the Lasso case), since the optimization problem might not be strictly convex. Consequently, there might exist other solutions that contain more than n active groups. However, a compact solution $\hat{\beta}$ with $|S(\hat{\beta})| \leq n$ can always be obtained by following an easy and mechanical step described in the proof of Proposition 2.2.

Proof: From the KKT condition in proposition 2.1, we know that any solution $\hat{\beta}$ should satisfy the following conditions ($j = 1, \dots, p_n$):

$$\frac{1}{n} X_j^T (Y - X\hat{\beta}) = \lambda g_j$$

where $g_j = \partial \|\beta_j\|_{\ell_q}$. Now suppose there is a solution $\hat{\beta}$ which has $s = |S(\hat{\beta})| > n$ number of active groups, in the following we will show that we can always construct another solution $\tilde{\beta}$ with one less active group, i.e. $|S(\tilde{\beta})| = |S(\hat{\beta})| - 1$.

Without loss of generality assume that the first s groups of variables in $\hat{\beta}$ are active, i.e. $\hat{\beta}_j \neq \mathbf{0}$ for $j = 1, \dots, s$. Since

$$X\hat{\beta} = \sum_{j=1}^s X_j \hat{\beta}_j \in \mathbb{R}^{n \times 1}$$

and $s > n$, the set of vectors $X_1 \hat{\beta}_1, \dots, X_s \hat{\beta}_s$ are linearly dependent. Without loss of generality assume

$$X_1 \hat{\beta}_1 = \alpha_2 X_2 \hat{\beta}_2 + \dots + \alpha_s X_s \hat{\beta}_s.$$

Now define $\tilde{\beta}_j = \mathbf{0}$ for $j = 1$ and $j > s$, and $\tilde{\beta}_j = (1 + \alpha_j) \hat{\beta}_j$ for $j = 2, \dots, s$, and it is straightforward to check that $\tilde{\beta}$ satisfies the KKT condition and thus is also a solution to the ℓ_1 - ℓ_q regularized regression problem in equation 2.1. The result thus follows by induction. \square

The main objective of the paper is to investigate several important statistical properties of the ℓ_1 - ℓ_q estimator $\hat{\beta}$. We first give some rough definitions of the properties that we would like to establish, more details will be shown in their corresponding sections.

Definition 2.4. (*Variable selection consistency*) An estimator is said to be variable selection consistent if it can correctly recover the sparsity pattern with probability goes to 1. For the case of grouped variable selection, $\hat{\beta}$ is said to be variable selection consistent if

$$\mathbb{P}\left(S(\hat{\beta}) = S(\beta^*)\right) \rightarrow 1. \quad (2.6)$$

Definition 2.5. (ℓ_1 -estimation consistency) An estimator is said to be ℓ_1 -estimation consistent if the ℓ_1 -norm of the difference between the estimator and the true parameter vector converges to 0 in probability. i.e.

$$\forall \delta > 0 \quad \mathbb{P}\left(\|\hat{\beta} - \beta^*\|_{\ell_1} > \delta\right) \rightarrow 0. \quad (2.7)$$

Definition 2.6. (*Prediction error consistency*) An estimator is said to be prediction error consistent if the prediction error, defined as $\frac{1}{n}\|\hat{Y} - X\beta^*\|_{\ell_2}^2$, of the estimator converges to 0 in probability. i.e.

$$\forall \delta > 0 \quad \mathbb{P}\left(\frac{1}{n}\|\hat{Y} - X\beta^*\|_{\ell_2}^2 > \delta\right) \rightarrow 0. \quad (2.8)$$

Definition 2.7. (*Risk consistency or Persistency*) Assuming the true model $f^*(X)$ does not have to be linear, for the regression model with random design, $(\mathcal{X}, \mathcal{Y}) \sim F_n \in \mathcal{F}^n$, where \mathcal{F}^n is a collection of distributions of i.i.d. $m_n + 1$ dimensional random vectors. Define the risk function under the distribution F_n to be $R_{F_n}(\beta)$ (More details in Section 5). Given a sequence of sets of predictors \mathcal{B}_n , the sequence of estimators $\hat{\beta}^{\hat{F}_n} \in \mathcal{B}_n$ is called persistent if for every sequence $F_n \in \mathcal{F}^n$,

$$R_{F_n}(\hat{\beta}^{\hat{F}_n}) - R_{F_n}(\beta_*^{F_n}) \xrightarrow{P} 0, \quad (2.9)$$

where

$$\beta_*^{F_n} = \arg \min_{\beta \in \mathcal{B}_n} R_{F_n}(\beta). \quad (2.10)$$

For the ℓ_1 - ℓ_q regularized regression, later, we will use $\mathcal{B}_n = \{\beta : \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_2}^2 \leq L_n\}$, for some $L_n = o((n/(\log n))^{1/4})$.

The following table gives a high level summary of our main results, ordered from very stringent assumptions to much weaker assumptions:

$$\text{Variable selection consistency:} \quad \mathbb{P}\left(S(\widehat{\beta}) = S(\beta^*)\right) \rightarrow 1 \quad (\text{R1})$$

$$\ell_1\text{-estimation convergence rate:} \quad \|\widehat{\beta} - \beta^*\|_{\ell_1} = O_P\left(s_n \bar{d}_n \sqrt{\frac{\log m_n}{n}}\right) \quad (\text{R2})$$

$$\text{Prediction error convergence rate:} \quad \frac{1}{n} \|\widehat{Y} - X\beta^*\|_{\ell_2}^2 = O_P\left(\frac{s_n \bar{d}_n \log m_n}{n}\right) \quad (\text{R3})$$

$$\text{Prediction (misspecified model):} \quad \frac{1}{n} \|\widehat{Y} - f^*\|_{\ell_2}^2 = O_P\left(\frac{s' \bar{d}_n \log m_n}{n}\right) \quad (\text{R3}^*)$$

$$\text{Persistency (misspecified model):} \quad R_{F_n}(\widehat{\beta}^{F_n}) - R_{F_n}(\beta_*^{F_n}) \xrightarrow{P} 0 \quad (\text{R4})$$

Remark 2.8. (R1) to (R3) assume the true model must be linear, while (R3*) and (R4) relax this condition so that the model can be misspecified. Even though (R3) and (R3*) look very similar, (R3*) dropped the linear model assumption at the price of enforcing another “*weak sparsity*” condition. Also, (R1), (R2), (R3), and (R3*) are fixed design results, while (R4) is a random design result.

In general, the condition for variable selection consistency is the strongest since it involves not only certain relations among n , λ_n , p_n , s_n , \bar{d}_n , but also the minimum absolute value of the parameters, $\rho_n^* = \min_{j \in S} \|\beta_j^*\|_\infty$. The ℓ_1 -estimation consistency and prediction error consistency requires weaker conditions than variable selection consistency. Unlike the previous properties, when the model is misspecified, the prediction error consistency in (R3*) follows from a sparse oracle inequality. Since both the sparsity oracle inequalities and persistency does not require the existence of a true linear model and thus is more general. Especially, the persistency is about the consistency of the predictive risk when considering random design and only need a very weak assumption about the design.

III. VARIABLE SELECTION CONSISTENCY

In this Section we study the conditions under which the ℓ_1 - ℓ_q estimator is variable selection consistent. Our proof is adapted from (Wainwright, 2006) and (Ravikumar et al., 2007). The former paper develop the “witness” proof idea which is the main framework used in our proof. The latter paper mainly treat variable selection consistency when $q = 2$ in a nonparametric sparse additive model setting, which makes their conditions more stringent than ours even when $q = 2$.

In the following, Let S denote the true set of group indices $\{j : X_j \neq 0\}$, with $s_n = |S|$, and S^c denote its complement. Denote $\Lambda_{\min}(C)$ to be the minimum eigenvalue of the matrix C .

Then, we have

Theorem 3.1. *Let q and q' are conjugate exponents with each other, that is $\frac{1}{q} + \frac{1}{q'} = 1$ and $1 \leq q, q' \leq \infty$. Suppose that the following conditions hold on the design matrix X :*

$$\Lambda_{\min} \left(\frac{1}{n} X_S^T X_S \right) \geq C_{\min} > 0$$

$$\max_{j \in S^c} \left\| (X_j^T X_S)(X_S^T X_S)^{-1} \right\|_{q', q'} \leq 1 - \delta, \quad \text{for some } 0 < \delta \leq 1. \quad (3.1)$$

where $\|\cdot\|_{a,b}$ is the matrix norm, defined as $\|A\|_{a,b} = \sup_x \frac{\|Ax\|_{\ell_b}}{\|x\|_{\ell_a}}$, $1 \leq a, b \leq \infty$. Assume the maximum number of variables with each group $\bar{d}_n \rightarrow \infty$ and $\bar{d}_n = o(n)$. Furthermore, suppose the following conditions, which relate the regularization parameter λ_n to the design parameters n , p_n , the number of relevant groups s_n and the maximum group size \bar{d}_n :

$$\frac{\lambda_n^2 n}{\log((p_n - s_n)\bar{d}_n)} \rightarrow \infty. \quad (3.2)$$

$$\frac{1}{\rho_n^*} \left\{ \sqrt{\frac{\log(s_n \bar{d}_n)}{n}} + \lambda_n (\bar{d}_n)^{1/q'} \left\| \left(\frac{1}{n} X_S^T X_S \right)^{-1} \right\|_{\infty, \infty} \right\} \rightarrow 0. \quad (3.3)$$

where $\rho_n^* = \min_{j \in S} \|\beta_j^*\|_{\infty}$. Then, the ℓ_1 - ℓ_q regularized regression is variable selection consistent.

Remark 3.2. First, notice that the result established in Theorem 3.1 is a direct generalization of the variable selection result for Lasso in (Wainwright, 2006) by setting $q = 1$ and $\bar{d}_n = 1$ (as then the ℓ_1 - ℓ_q degenerates to Lasso). This gives the sufficient conditions for exact recovery of sparsity pattern in β^* for the ℓ_1 - ℓ_q regularized regression. Also notice that when \bar{d}_n is bounded from above, the conditions are almost the same as those of Lasso except the condition in equation 3.1 which depends on the value of q .

Second, we consider the case when ρ_n is bounded away from zero. Assuming that $q = \infty$ and $\bar{d}_n = n^{1/5}$ (such as in the fitting of additive model with basis expansion), we must have $\lambda_n = o(n^{-1/5})$ and as a result of $\frac{\lambda_n^2 n}{\log((p_n - s_n)\bar{d}_n)} \rightarrow \infty$, we need to have $p_n = o(\exp(n^{3/5}))$. This means that even when we have increasing group size \bar{d}_n , the sparse pattern (in terms of grouped variables) can still be correctly identified with a large p_n .

Finally, when minimum parameter value $\rho_n \rightarrow 0$, to ensure variable selection consistency, it can at most converge to zero at a rate slower than $n^{-1/2}$.

Proof: Note, the special case when $q = 1$ has already been proved in (Wainwright et al., 2006). Here, we only consider the case that $1 < q \leq \infty$. A vector $\hat{\beta} \in \mathbb{R}^{m_n}$, $m_n = \sum_{j=1}^{p_n} d_j$,

is an optimum of the objective function in (2.1) if and only if there exists a sequence of subgradients $\widehat{g}_j \in \partial \|\widehat{\beta}_j\|_{\ell_q}$, such that

$$\frac{1}{n} X^T \left(\sum_j X_j \widehat{\beta}_j - Y \right) + \lambda_n (d_j)^{1/q'} \widehat{g}_j = \mathbf{0}. \quad (3.4)$$

The subdifferentials $\partial \|\widehat{\beta}_j\|_{\ell_q}$ satisfies the KKT conditions in proposition 2.1.

Our argument closely follows the approach of Wainwright et al. (2006) in the linear case. In particular, we proceed by a “witness” proof technique, to show the existence of a coefficient-subgradient pair $(\widehat{\beta}, \widehat{g})$ for which $\text{supp}(\widehat{\beta}) = \text{supp}(\beta^*)$. To do so, we first set $\widehat{\beta}_{S^c} = \mathbf{0}$ and \widehat{g}_S to be the vector concatenating all the subvectors \widehat{g}_j 's, for $j \in S$. We also define \widehat{g}_{S^c} and $\widehat{\beta}_S$ in a similar way. And we then obtain $\widehat{\beta}_S$ and \widehat{g}_{S^c} from the stationary conditions in (3.4). By showing that, with high probability,

$$\widehat{\beta}_j \neq \mathbf{0} \text{ for } j \in S \quad (3.5)$$

$$\widehat{g}_j \in B^{q'}(1) \text{ for } j \in S^c, \quad (3.6)$$

this demonstrates that with high probability there exists an optimal solution to the optimization problem in (2.1) that has the same sparsity pattern as the true model.

Setting $\widehat{\beta}_{S^c} = \mathbf{0}$ and

$$\widehat{g}_j = \begin{cases} \left\{ \left(\frac{|\widehat{\beta}_{j\ell}|^{q-1} \text{sign}(\widehat{\beta}_{j\ell})}{\|\widehat{\beta}_j\|_{\ell_q}^{q-1}} \right)_{\ell=1}^{d_j} \right\} & 1 < q < \infty \\ \text{conv}\{\text{sign}(\widehat{\beta}_{j\ell}) e_\ell : |\widehat{\beta}_{j\ell}| = \|\widehat{\beta}_j\|_{\ell_\infty}\} & q = \infty \end{cases} \quad (3.7)$$

for $j \in S$, denote $W = \text{diag}((d_1)^{1/q'} I_{d_1}, \dots, (d_p)^{1/q'} I_{d_p})$ where I_{d_j} is a $d_j \times d_j$ identity matrix. We define W_S to be submatrix of W by extracting out the rows and columns corresponding to the group index set S . The stationary condition for $\widehat{\beta}_S$ is

$$\frac{1}{n} X_S^T (X_S \widehat{\beta}_S - Y) + \lambda_n W_S \widehat{g}_S = \mathbf{0}. \quad (3.8)$$

Let $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$, then the stationary condition can be written as

$$\frac{1}{n} X_S^T X_S (\widehat{\beta}_S - \beta_S^*) - \frac{1}{n} X_S^T \epsilon + \lambda_n W_S \widehat{g}_S = \mathbf{0} \quad (3.9)$$

or

$$\widehat{\beta}_S - \beta_S^* = \left(\frac{1}{n} X_S^T X_S \right)^{-1} \left(\frac{1}{n} X_S^T \epsilon - \lambda_n W_S \widehat{g}_S \right) \quad (3.10)$$

assuming that $\frac{1}{n} X_S^T X_S$ is nonsingular. Recalling our definition

$$\rho_n^* = \min_{j \in S} \|\beta_j^*\|_{\ell_\infty} > 0. \quad (3.11)$$

it suffices to show that

$$\|\widehat{\beta}_S - \beta_S^*\|_{\ell_\infty} < \frac{\rho_n^*}{2} \quad (3.12)$$

in order to ensure that $\text{supp}(\beta_S^*) = \text{supp}(\widehat{\beta}_S) = \{j : \|\widehat{\beta}_j\|_{\ell_\infty} \neq 0\}$.

Using $\Sigma_{SS} = \frac{1}{n}X_S^T X_S$ to simplify notation, we have the ℓ_∞ bound

$$\|\widehat{\beta}_S - \beta_S^*\|_{\ell_\infty} \leq \left\| \Sigma_{SS}^{-1} \left(\frac{1}{n} X_S^T \epsilon \right) \right\|_{\ell_\infty} + \lambda_n \|\Sigma_{SS}^{-1} W_S \widehat{g}_S\|_{\ell_\infty}. \quad (3.13)$$

We now proceed to bound the quantities above. First note that for $j \in S$, $\|\widehat{g}_j\|_{\ell_{q'}} = 1$. Therefore, since

$$\|\widehat{g}_S\|_{\ell_\infty} = \max_{j \in S} \|\widehat{g}_j\|_{\ell_\infty} \leq \max_{j \in S} \|\widehat{g}_j\|_{\ell_{q'}} = 1 \quad (3.14)$$

we have that

$$\|\Sigma_{SS}^{-1} W_S \widehat{g}_S\|_{\ell_\infty} \leq (\bar{d}_n)^{1/q'} \|\Sigma_{SS}^{-1}\|_{\infty, \infty}. \quad (3.15)$$

Therefore

$$\|\widehat{\beta}_S - \beta_S^*\|_{\ell_\infty} \leq \left\| \Sigma_{SS}^{-1} \left(\frac{1}{n} X_S^T \epsilon \right) \right\|_{\ell_\infty} + \lambda_n (\bar{d}_n)^{1/q'} \|\Sigma_{SS}^{-1}\|_{\infty, \infty}.$$

Finally, consider $Z = \Sigma_{SS}^{-1} \left(\frac{1}{n} X_S^T \epsilon \right)$. Note that $\epsilon \sim N(0, \sigma^2 I)$, so that Z is Gaussian as well, with mean zero. Consider its ℓ -th component, $Z_\ell = e_\ell^T Z$. Then $\mathbb{E}[Z_\ell] = 0$, and

$$\text{Var}(Z_\ell) = \frac{\sigma^2}{n} e_\ell^T \Sigma_{SS}^{-1} e_\ell \leq \frac{\sigma^2}{nC_{\min}}. \quad (3.16)$$

By the comparison results on Gaussian maxima (Ledoux and Talagrand, 1991), we have then that

$$\mathbb{E}[\|Z\|_{\ell_\infty}] \leq 3\sqrt{\log(s\bar{d}_n)} \max_{\underline{\ell}} \sqrt{\text{Var}(Z_{\underline{\ell}})} \leq 3\sigma \sqrt{\frac{\log(s\bar{d}_n)}{nC_{\min}}}. \quad (3.17)$$

An application of Markov's inequality then gives that

$$\begin{aligned} \mathbb{P}\left(\|\widehat{\beta}_S - \beta_S^*\|_{\ell_\infty} > \frac{\rho_n^*}{2}\right) &\leq \mathbb{P}\left(\|Z\|_{\ell_\infty} + \lambda_n (\bar{d}_n)^{1/q'} \|\Sigma_{SS}^{-1}\|_{\infty, \infty} > \frac{\rho_n^*}{2}\right) \\ &\leq \frac{2}{\rho_n^*} \left\{ \mathbb{E}[\|Z\|_{\ell_\infty}] + \lambda_n (\bar{d}_n)^{1/q'} \|\Sigma_{SS}^{-1}\|_{\infty, \infty} \right\} \end{aligned} \quad (3.18)$$

$$\leq \frac{2}{\rho_n^*} \left\{ 3\sigma \sqrt{\frac{\log(s\bar{d}_n)}{nC_{\min}}} + \lambda_n (\bar{d}_n)^{1/q'} \|\Sigma_{SS}^{-1}\|_{\infty, \infty} \right\} \quad (3.19)$$

which converges to zero under the condition that

$$\frac{1}{\rho_n^*} \left\{ \sqrt{\frac{\log(s\bar{d}_n)}{n}} + \lambda_n (\bar{d}_n)^{1/q'} \|\Sigma_{SS}^{-1}\|_{\infty, \infty} \right\} \longrightarrow 0. \quad (3.20)$$

We now analyze \widehat{g}_{S^c} . Recall that we have set $\widehat{\beta}_{S^c} = \beta_{S^c}^* = 0$. The stationary condition for $j \in S^c$ is thus given by

$$\frac{1}{n} X_j^T \left(X_S \widehat{\beta}_S - X_S \beta_S^* - \epsilon \right) + \lambda_n (d_j)^{1/q'} \widehat{g}_j = \mathbf{0}. \quad (3.21)$$

Therefore,

$$\begin{aligned} \widehat{g}_{S^c} &= \frac{W_{S^c}^{-1}}{\lambda_n} \left\{ \frac{1}{n} X_{S^c}^T X_S \left(\beta_S^* - \widehat{\beta}_S \right) + \frac{1}{n} X_{S^c}^T \epsilon \right\} \\ &= \frac{W_{S^c}^{-1}}{\lambda_n} \left\{ \frac{1}{n} X_{S^c}^T X_S \left(\frac{1}{n} X_S^T X_S \right)^{-1} \left(\lambda_n W_S \widehat{g}_S - \frac{1}{n} X_S^T \epsilon \right) + \frac{1}{n} X_{S^c}^T \epsilon \right\} \\ &= \frac{W_{S^c}^{-1}}{\lambda_n} \left\{ \Sigma_{S^c S} \Sigma_{SS}^{-1} \left(\lambda_n W_S \widehat{g}_S - \frac{1}{n} X_S^T \epsilon \right) + \frac{1}{n} X_{S^c}^T \epsilon \right\} \end{aligned} \quad (3.22)$$

from equation (3.10).

We want to show that

$$\widehat{g}_j \in B^{q'}(1) \quad (3.23)$$

for all $j \in S^c$. From (3.22), we see that \widehat{g}_j is Gaussian, with mean

$$\mu_j = \mathbb{E}(\widehat{g}_j) = (d_j)^{-1/q'} \Sigma_{jS} \Sigma_{SS}^{-1} W_S \widehat{g}_S. \quad (3.24)$$

We then obtain the bound

$$\|\mu_j\|_{\ell_{q'}} \leq \|\Sigma_{jS} \Sigma_{SS}^{-1}\|_{q', q'} \|\widehat{g}_S\|_{\ell_{q'}} = \|\Sigma_{jS} \Sigma_{SS}^{-1}\|_{q', q'} \leq 1 - \delta \quad \text{for some } \delta > 0.$$

It therefore suffices to show that

$$\mathbb{P} \left(\max_{j \in S^c} (d_j)^{1/q'} \|\widehat{g}_j - \mu_j\|_{\ell_\infty} > \frac{\delta}{2} \right) \longrightarrow 0 \quad (3.25)$$

since this implies that

$$\|\widehat{g}_j\|_{\ell_{q'}} \leq \|\mu_j\|_{\ell_{q'}} + \|\widehat{g}_j - \mu_j\|_{\ell_{q'}} \quad (3.26)$$

$$\leq \|\mu_j\|_{\ell_{q'}} + (d_j)^{1/q'} \|\widehat{g}_j - \mu_j\|_{\ell_\infty} \quad (3.27)$$

$$\leq (1 - \delta) + \frac{\delta}{2} + o(1) \quad (3.28)$$

with probability approaching one. To show (3.25), we again appeal to comparison results of Gaussian maxima. Define

$$Z_j = (d_j)^{1/q'} \lambda_n (\widehat{g}_j - \mu_j) = X_j^T (I - X_S (X_S^T X_S)^{-1} X_S^T) \frac{\epsilon}{n} \quad (3.29)$$

for $j \in S^c$. Then Z_j are zero mean Gaussian random vector, and we need to show that

$$\mathbb{P} \left(\max_{j \in S^c} \frac{\|Z_j\|_{\ell_\infty}}{\lambda_n} \geq \frac{\delta}{2} \right) \longrightarrow 0. \quad (3.30)$$

Let Z_{jk} represent the k -th element of Z_j for $j \in S^c$. A calculation shows that $\mathbb{E}(Z_{jk}^2) \leq \frac{\sigma^2}{n}$. Therefore, we have by Markov's inequality and the comparison results of Gaussian maxima that

$$\begin{aligned} \mathbb{P}\left(\max_{j \in S^c} \frac{\|Z_j\|_{\ell_\infty}}{\lambda_n} \geq \frac{\delta}{2}\right) &\leq \frac{2}{\delta \lambda_n} \mathbb{E}\left(\max_{j \in S^c, k} |Z_{jk}|\right) \\ &\leq \frac{2}{\delta \lambda_n} \left(3\sqrt{\log((p_n - s_n)\bar{d}_n)} \max_{j \in S^c, k} \sqrt{\mathbb{E}(Z_{jk}^2)}\right) \end{aligned} \quad (3.31)$$

$$\leq \frac{6\sigma}{\delta \lambda_n} \sqrt{\frac{\log((p_n - s_n)\bar{d}_n)}{n}} \quad (3.32)$$

which converges to zero under the condition that

$$\frac{\lambda_n^2 n}{\log((p_n - s_n)\bar{d}_n)} \longrightarrow \infty. \quad (3.33)$$

This is just the condition in the statement of the theorem. \square

IV. ESTIMATION CONSISTENCY

In this section, we prove the estimation consistency results under two types assumptions:

- (i) When the model is correctly specified, i.e., the true model is linear, we can achieve both ℓ_1 -consistency results and derive the optimal rate of convergence for the prediction error.
- (ii) When the model is misspecified, i.e. the true model is not linear, we can still achieve a sparsity oracle inequality, which provide a bound of the prediction error using the loss of the prediction oracle with the number of nonzero groups of the prediction loss involved in. Under the “*weak sparsity*” condition, we can still obtain a rate of convergence of the prediction error which is similar to the convergence rate obtained under the linear model assumption.

We begin with a technical lemma, which is essentially lemma 1 as in (Bunea et al., 2007a) and (Bickel et al., 2007), but need to be extended to handle the group structures in the more general ℓ_1 - ℓ_q regularized regression setting.

Lemma 4.1. *Let $\epsilon_1, \dots, \epsilon_n$ be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$ and Let $\hat{Y} = X\hat{\beta}$ be the ℓ_1 - ℓ_q regularized regression estimator with $1 \leq q \leq \infty$ as in (2.1) with*

$$\lambda_n = A\sigma \sqrt{\frac{\log m_n}{n}} \quad (4.1)$$

for some $A > 2\sqrt{2}$. Then, for all $m_n \geq 2$, $n > 1$, with probability of at least $1 - m_n^{1-A^2/8}$ we have simultaneously for all $\beta \in \mathbb{R}^{m_n}$:

$$\frac{1}{n} \|\hat{Y} - X\beta^*\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j\|_{\ell_q} \leq \frac{1}{n} \|X\beta - X\beta^*\|_{\ell_2}^2 + 4 \sum_{j \in S(\beta)} \lambda_n (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j\|_{\ell_q} \quad (4.2)$$

where $S(\beta)$ denotes the set of nonzero group indices of β .

Proof: By the definition of $\widehat{Y} = X\widehat{\beta}$, we have

$$\frac{1}{2n}\|Y - X\widehat{\beta}\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j\|_{\ell_q} \leq \frac{1}{2n}\|Y - X\beta\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q}$$

for all $\beta \in \mathbb{R}^{m_n}$, $m_n = \sum_{j=1}^{p_n} d_j$, which we may rewritten as

$$\begin{aligned} & \frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_{\ell_2}^2 + 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j\|_{\ell_q} \\ & \leq \frac{1}{n}\|X\beta^* - X\beta\|_{\ell_2}^2 + 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q} + \frac{2}{n}\epsilon^T X(\widehat{\beta} - \beta). \end{aligned} \quad (4.3)$$

For each $\underline{j} = 1, \dots, m_n$, we define the random variables $V_{\underline{j}} = \frac{1}{n}X_{\underline{j}}^T \epsilon$, and the event

$$\mathcal{A} = \bigcap_{\underline{j}=1}^{m_n} \left\{ 2|V_{\underline{j}}| \leq \lambda_n \right\}.$$

Under the normality assumption, we have that

$$\sqrt{n}V_{\underline{j}} \sim \mathcal{N}(0, \sigma^2) \quad \underline{j} = 1, \dots, m_n. \quad (4.4)$$

Using the elementary bound on the tails of Gaussian distribution we find that the probability of the complementary event \mathcal{A}^c satisfies

$$\mathbb{P}\{\mathcal{A}^c\} \leq \sum_{\underline{j}=1}^{m_n} \mathbb{P}\{\sqrt{n}|V_{\underline{j}}| > \sqrt{n}\lambda_n/2\} \leq m_n \mathbb{P}\{|Z| \geq \sqrt{n}\lambda_n/(2\sigma)\} \quad (4.5)$$

$$\leq m_n \exp\left(-\frac{n\lambda_n^2}{8\sigma^2}\right) = m_n \exp\left(-\frac{A^2 \log m_n}{8}\right) = m_n^{1-A^2/8} \quad (4.6)$$

where $Z \sim \mathcal{N}(0, 1)$. Then, on the set \mathcal{A} , we have

$$\frac{2}{n}\epsilon^T X(\widehat{\beta} - \beta) = 2 \sum_{\underline{j}=1}^{m_n} V_{\underline{j}}(\widehat{\beta}_{\underline{j}} - \beta_{\underline{j}}) \leq \sum_{\underline{j}=1}^{m_n} \lambda_n |\widehat{\beta}_{\underline{j}} - \beta_{\underline{j}}| \leq \sum_{j=1}^{p_n} \lambda_n (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q}$$

and therefore, still on the set \mathcal{A} ,

$$\begin{aligned} & \frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_{\ell_2}^2 \leq \frac{1}{n}\|X\beta^* - X\beta\|_{\ell_2}^2 \\ & + 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q} + \sum_{j=1}^{p_n} \lambda_n (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} - 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j\|_{\ell_q}. \end{aligned} \quad (4.7)$$

Adding the same term $\sum_{j=1}^{p_n} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q}$ on both sides, we obtain

$$\begin{aligned} \frac{1}{n} \|X\beta^* - X\widehat{\beta}\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} &\leq \frac{1}{n} \|X\beta^* - X\beta\|_{\ell_2}^2 \\ &+ 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q} + 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} - 2\lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j\|_{\ell_q}. \end{aligned} \quad (4.8)$$

Recall $S(\beta)$ to be the set of non-zero group indices of β . Rewriting the right-hand side of the previous display, then, on set \mathcal{A}

$$\begin{aligned} \frac{1}{n} \|X\beta^* - X\widehat{\beta}\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \\ \leq \frac{1}{n} \|X\beta^* - X\beta\|_{\ell_2}^2 + 2 \left(\sum_{j=1}^{p_n} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} - \sum_{j \notin S(\beta)} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j\|_{\ell_q} \right) \\ + 2 \left(\sum_{j \in S(\beta)} \lambda_n(d_j)^{1/q'} \|\beta_j\|_{\ell_q} - \sum_{j \in S(\beta)} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j\|_{\ell_q} \right) \\ \leq \frac{1}{n} \|X\beta - X\beta^*\|_{\ell_2}^2 + 4 \sum_{j \in S(\beta)} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \end{aligned}$$

by the triangle inequality and the fact that $\beta_j = \mathbf{0}$ for $j \notin S(\beta)$. \square

A. Estimation Consistency Under the Linear Model Assumption

Assuming the true model is linear, to obtain the ℓ_1 -consistency result, a key assumption on the design matrix is needed, which is stated as the following

Assumption 1 Recall that $s_n = S(\beta^*)$, assume for any vector $\gamma \in \mathbb{R}^{m_n}$ satisfies

$$\kappa \equiv \min_{S_0 \subset \{1, \dots, p\}: |S_0| \leq s_n} \min_{\sum_{j \in S_0^c} (d_j)^{1/q'} \|\gamma_j\|_{\ell_q} \leq 3 \sum_{j \in S_0} (d_j)^{1/q'} \|\gamma_j\|_{\ell_q}} \frac{\|X\gamma\|_{\ell_2}}{\sqrt{n} \sqrt{\sum_{j \in S_0} (d_j)^{2/q'-1} \|\gamma_j\|_{\ell_q}^2}} > 0 \quad (4.9)$$

Remark 4.2. Before proving the following theorem, we pause to make some comments about this assumption.

First, For $q = 1$ (thus, $q' = \infty$), this assumption is very similar to the *restricted eigenvalue* assumption as in (Bickel et al., 2007), which is defined as

$$\kappa \equiv \min_{S_0 \subset \{1, \dots, p\}: |S_0| \leq s_n} \min_{\sum_{j \in S_0^c} \|\gamma_j\|_{\ell_1} \leq 3 \sum_{j \in S_0} \|\gamma_j\|_{\ell_1}} \frac{\|X\gamma\|_{\ell_2}}{\sqrt{n} \sqrt{\sum_{j \in S_0} \|\gamma_j\|_{\ell_2}^2}} > 0. \quad (4.10)$$

However, our assumption is slightly weaker, due to the fact that, for any $\gamma \in \mathbb{R}^{d_j}$

$$\|\gamma_j\|_{\ell_1}^2 \leq d_j \|\gamma_j\|_{\ell_2}^2. \quad (4.11)$$

Second, the quantity $\sqrt{\sum_{j \in S_0} (d_j)^{2/q'-1} \|\gamma_j\|_{\ell_q}^2}$ in our assumption balances between $q = 1$ and $q = \infty$. For example, when $q = 1$, $\|\gamma_j\|_{\ell_1}^2$ is relatively large, but $(d_j)^{2/q'-1} = (d_j)^{-1}$ is very small. While for $q = \infty$, $\|\gamma_j\|_{\ell_q}^2 = \|\gamma_j\|_{\ell_\infty}^2$ is relatively small, however, $(d_j)^{2/q'-1} = (d_j)^1$ is very significant. In this sense, $q = 2$ seems the most balanced one, due to the fact that

$$\sum_{j \in S_0} (d_j)^{-1} \|\gamma_j\|_{\ell_1}^2 \leq \sum_{j \in S_0} \sqrt{d_j} \|\gamma_j\|_{\ell_2}^2 \leq \sum_{j \in S_0} d_j \|\gamma_j\|_{\ell_\infty}^2 \quad (4.12)$$

Therefore, among $q = 1, 2, \infty$, $q = 2$ needs the weakest assumption, this provides some insights about why group Lasso might also be a suitable choice for grouped variable selection. However, we need to more cautions to say which value of q is the best. Since in real applications, the choice of q might depends on the true relevant coefficients β_S^* . If different components in the relevant groups are on the same order of magnitude, $q = \infty$ might be more suitable, on the contrary, if some relevant coefficients are very small relative to the others, $q = 1$ might be better. we plan to investigate this issue in a separate paper.

Theorem 4.3. (*Estimation consistency under linear model assumptions*) Under assumption 2, let $\epsilon_1, \dots, \epsilon_n$ be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Consider the ℓ_1 - ℓ_q regularized estimator defined by (2.1) with

$$\lambda_n = A\sigma \sqrt{\frac{\log m_n}{n}} \quad (4.13)$$

for some $A > 2\sqrt{2}$. then, for all $n \geq 1$ with probability at least $1 - m_n^{1-A^2/8}$ we have

$$\frac{1}{n} \|\widehat{Y} - X\beta^*\|_{\ell_2}^2 \leq \frac{9A^2\sigma^2}{\kappa^2} \frac{s_n \bar{d}_n \log m_n}{n} \quad (4.14)$$

$$\|\widehat{\beta} - \beta^*\|_{\ell_1} \leq \frac{12A^2\sigma^2 s_n \bar{d}_n}{\kappa^2} \sqrt{\frac{\log m_n}{n}}. \quad (4.15)$$

Remark 4.4. From this theorem, we obtain ℓ_1 -consistency and the corresponding rate of convergence. Due to the fact that $\|\gamma\|_{\ell_q} \leq \|\gamma\|_{\ell_1}$ for all $1 < q \leq \infty$, we obtain ℓ_q consistency also if $s_n \bar{d}_n \sqrt{\frac{\log m_n}{n}} \rightarrow 0$. If we want to the rate of convergence for ℓ_2 -consistency, a direct result will be

$$\|\widehat{\beta} - \beta^*\|_{\ell_2}^2 \leq \frac{144A^4\sigma^4 s_n^2 \bar{d}_n^2 \log m_n}{\kappa^4 n}. \quad (4.16)$$

which is suboptimal. Recall that $\|\widehat{\beta} - \beta^*\|_{\ell_1}^2 \leq p_n \bar{d}_n \|\widehat{\beta} - \beta^*\|_{\ell_2}^2$, if $|S(\widehat{\beta})|$ is $O(s_n)$ and the elements in $\widehat{\beta}_j - \beta_j^*$ are balanced for $j \in S$, then we can also achieve the optimal rate of convergence for ℓ_2 -norm consistency. How to obtain optimal rate of convergence for ℓ_q -consistency for general q would be an interesting future work.

Proof: From equation 4.2, Using $\beta = \beta^*$, we have that on the event \mathcal{A} ,

$$\frac{1}{n} \|\hat{Y} - X\beta^*\|_{\ell_2}^2 \leq 3 \sum_{j \in S(\beta^*)} \lambda_n(d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} \leq 3\lambda_n \sqrt{\bar{d}_n s_n} \sqrt{\sum_{j \in S(\beta^*)} (d_j)^{2/q'-1} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q}^2} \quad (4.17)$$

$$\sum_{j \in S(\beta^*)^c} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} \leq 3 \sum_{j \in S(\beta^*)} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} \quad (4.18)$$

By the last equation, we have that assumption 1 hold on event \mathcal{A} , by this assumption, we have that

$$\frac{1}{n} \|\hat{Y} - X\beta^*\|_{\ell_2}^2 \geq \kappa^2 \sum_{j \in S(\beta^*)} (d_j)^{2/q'-1} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q}^2. \quad (4.19)$$

By combining the above inequalities, we get

$$\frac{1}{n} \|\hat{Y} - X\beta^*\|_{\ell_2}^2 \leq \frac{9\lambda_n^2 s_n \bar{d}_n}{\kappa^2} \quad (4.20)$$

and

$$\sqrt{\sum_{j \in S(\beta^*)} (d_j)^{2/q'-1} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q}^2} \leq \frac{3\lambda_n \sqrt{\bar{d}_n s_n}}{\kappa^2}. \quad (4.21)$$

Thus, we have

$$\|\hat{\beta} - \beta^*\|_{\ell_1} = \sum_{j=1}^{p_n} \|\hat{\beta}_j - \beta_j^*\|_{\ell_1} \leq \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} \quad (4.22)$$

$$= \sum_{j \in S(\beta^*)} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} + \sum_{j \in S(\beta^*)^c} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} \quad (4.23)$$

$$\leq 4 \sum_{j \in S(\beta^*)} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q} \leq 4\sqrt{\bar{d}_n s_n} \sqrt{\sum_{j \in S(\beta^*)} (d_j)^{2/q'-1} \|\hat{\beta}_j - \beta_j^*\|_{\ell_q}^2} \quad (4.24)$$

$$\leq \frac{12\lambda_n \bar{d}_n s_n}{\kappa^2} = \frac{12A^2 \sigma^2 s_n \bar{d}_n}{\kappa^2} \sqrt{\frac{\log m_n}{n}}. \quad (4.25)$$

Note, equation 4.20 is exactly equation 4.17. \square

B. Oracle Inequalities for Prediction Error Under Misspecified Models

Assuming the true regression function $f^*(X)$ is not linear, i.e. the model is misspecified. We can no longer obtain the optimal rate of convergence directly. But we can still obtain a sparsity oracle inequality, which can bound the prediction error in terms of nonzero components of the prediction oracle.

Assumption 2 Assume s' is an integer such that $1 \leq s' \leq p_n$, and δ is some positive number, then, for any $\gamma \neq 0$

$$\kappa(s', \delta) \equiv \min_{S_0 \subset \{1, \dots, p\}: |S_0| \leq s'} \min_{\sum_{j \in S_0^c} (d_j)^{1/q'} \|\gamma_j\|_{\ell_q} \leq (2 + \frac{3}{\delta}) \sum_{j \in S_0} (d_j)^{1/q'} \|\gamma_j\|_{\ell_q}} \frac{\|X\gamma\|_{\ell_2}}{\sqrt{n} \sqrt{\sum_{j \in S_0} (d_j)^{2/q'-1} \|\gamma_j\|_{\ell_q}^2}} > 0.$$

Theorem 4.5. Under assumption (2), let $\epsilon_1, \dots, \epsilon_n$ be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Consider the ℓ_1 - ℓ_q regularized estimator defined by (2.1) with

$$\lambda_n = A\sigma \sqrt{\frac{\log m_n}{n}} \quad (4.26)$$

for some $A > 2\sqrt{2}$. then, for all $n \geq 1$ with probability at least $1 - m_n^{1-A^2/8}$ we have

$$\begin{aligned} & \frac{1}{n} \|f^* - X\hat{\beta}\|_{\ell_2}^2 \\ & \leq (1 + \delta) \inf_{\beta \in \mathbb{R}^{m_n}: |S(\beta)| \leq s'} \left\{ \frac{1}{n} \|f^* - X\beta\|_{\ell_2}^2 + \frac{C(\delta)A^2\sigma^2}{\kappa(s', \delta)^2} \left(\frac{\bar{d}_n |S(\beta)| \log m_n}{n} \right) \right\} \end{aligned} \quad (4.27)$$

where $C(\delta) > 0$ is a constant depending only on δ . While $|S(\beta)|$ represents the number of nonzero elements in the set $S(\beta)$.

Remark 4.6. From this sparsity oracle inequality, if we add some assumptions, such as there exists some β' , such that $\frac{1}{n} \|f^* - X\beta'\|_{\ell_2}^2 \rightarrow 0$, then we can still obtain prediction error consistency if $\frac{\bar{d}_n |S(\beta')| \log m_n}{n} \rightarrow 0$. If we also want to obtain a convergence rate similar to that as in theorem 4.3, more conditions will be needed, as is shown in corollary 4.8.

Proof: Fix an arbitrary $\beta \in \mathbb{R}^{m_n}$ with $|S(\beta)| \leq s'$. On the event \mathcal{A} , we get from lemma 4.1 that

$$\begin{aligned} & \frac{1}{n} \|\hat{Y} - f^*\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j\|_{\ell_q} \\ & \leq \frac{1}{n} \|X\beta - f^*\|_{\ell_2}^2 + 4 \sum_{j \in S(\beta)} \lambda_n (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j\|_{\ell_q} \end{aligned} \quad (4.28)$$

Further from above, we can get that

$$\frac{1}{n} \|\hat{Y} - f^*\|_{\ell_2}^2 \leq \frac{1}{n} \|X\beta - f^*\|_{\ell_2}^2 + 3\lambda_n \sum_{j \in S(\beta)} (d_j)^{1/q'} \|\hat{\beta}_j - \beta_j\|_{\ell_q} \quad (4.29)$$

$$\leq \frac{1}{n} \|X\beta - f^*\|_{\ell_2}^2 + 3\lambda_n \sqrt{\bar{d}_n |S(\beta)|} \sqrt{\sum_{j \in S(\beta)} (d_j)^{2/q'-1} \|\hat{\beta}_j - \beta_j\|_{\ell_q}^2} \quad (4.30)$$

Consider separately the cases where

$$3 \sum_{j \in S(\beta)} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \leq \frac{\delta}{n} \|X\beta - f^*\|_{\ell_2}^2 \quad (4.31)$$

and

$$3 \sum_{j \in S(\beta)} \lambda_n(d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} > \frac{\delta}{n} \|X\beta - f^*\|_{\ell_2}^2 \quad (4.32)$$

In case (4.31), the result of the theorem trivially follows from equation (4.28). So, we will only consider the case (4.32). All the subsequent inequalities are valid on the event $\mathcal{A} \cap \mathcal{A}_1$ where \mathcal{A}_1 is defined by (4.32). On this event, we get from (4.28) that

$$\sum_{j=1}^{p_n} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \leq 3 \left(1 + \frac{1}{\delta}\right) \sum_{j \in S(\beta)} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \quad (4.33)$$

which further implies that

$$\sum_{j \in S(\beta)^c} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \leq \left(2 + \frac{3}{\delta}\right) \sum_{j \in S(\beta)} (d_j)^{1/q'} \|\widehat{\beta}_j - \beta_j\|_{\ell_q} \quad (4.34)$$

By assumption 2, we have

$$\kappa(s', \delta) \sqrt{\sum_{j \in S(\beta)} (d_j)^{2/q'-1} \|\widehat{\beta}_j - \beta_j\|_{\ell_q}^2} \leq \sqrt{\frac{1}{n} \|X(\widehat{\beta} - \beta)\|_{\ell_2}^2} = \frac{1}{\sqrt{n}} \|\widehat{Y} - X\beta\|_{\ell_2} \quad (4.35)$$

Combining this with (4.30), we get

$$\frac{1}{n} \|\widehat{Y} - f^*\|_{\ell_2}^2 \leq \frac{1}{n} \|X\beta - f^*\|_{\ell_2}^2 + 3\lambda_n \kappa^{-1}(s', \delta) \sqrt{\bar{d}_n |S(\beta)|} \left(\frac{1}{\sqrt{n}} \|\widehat{Y} - X\beta\|_{\ell_2} \right) \quad (4.36)$$

$$\begin{aligned} &\leq \frac{1}{n} \|X\beta - f^*\|_{\ell_2}^2 + 4\lambda_n \kappa^{-1}(s', \delta) \sqrt{\bar{d}_n |S(\beta)|} \left(\frac{1}{\sqrt{n}} \|\widehat{Y} - f^*\|_{\ell_2} \right. \\ &\quad \left. + \frac{1}{\sqrt{n}} \|X\beta - f^*\|_{\ell_2} \right) \end{aligned} \quad (4.37)$$

This inequality is of the same form as (A.4) in (Bunea et al., 2007a). A standard decoupling argument as in (Bunea et al., 2007a) using inequality $2xy \leq \frac{x^2}{b} + by^2$ with $b > 1$, $x = \lambda_n \kappa^{-1}(s', \delta) \sqrt{\bar{d}_n |S(\beta)|}$, and y being either $\frac{1}{\sqrt{n}} \|\widehat{Y} - f^*\|_{\ell_2}$ or $\frac{1}{\sqrt{n}} \|X\beta - f^*\|_{\ell_2}$ yields that

$$\frac{1}{n} \|\widehat{Y} - f^*\|_{\ell_2}^2 \leq \frac{b+1}{b-1} \frac{1}{n} \|X\beta - f^*\|_{\ell_2}^2 + \frac{8b^2}{(b-1)\kappa^2(s', \delta)} \lambda_n^2 \bar{d}_n |S(\beta)|, \quad \forall \beta > 1. \quad (4.38)$$

Taking $b = 1 + 2/\delta$ in the last display finishes the proof of the theorem. \square

From the above sparse oracle inequalities, we can show that the ℓ_1 - ℓ_q regression estimator can achieve the optimal rate of convergence if some “*weak sparsity*” condition holds (Bunea et al., 2007b). The main intuition is, even if the true function f^* can not be represented exactly by a linear model $X\beta$, but for some $\tilde{\beta} \in \mathbb{R}^{m_n}$ the squared distance from f^* to $X\beta$ can be controlled, up to logarithmic factors, by $|S(\tilde{\beta})|/n$. Then, the optimal rate of convergence can still be achieved. More formally, we define an oracle set as

Definition 4.7. Let B be a constant depending only for f^* and define an oracle set as

$$\mathcal{B} = \left\{ \beta : \text{ s.t. } \frac{1}{n} \|f^* - X\beta\|_{\ell_2}^2 \leq B\lambda_n^2 |S(\beta)| \right\} \quad (4.39)$$

Corollary 4.8. Under the same condition as in theorem 4.5, if the oracle set \mathcal{B} is nonempty and there is at least one element $\tilde{\beta}$ such that $|S(\tilde{\beta})| \leq s'$, we have

$$\frac{1}{n} \|f^* - \hat{Y}\|_{\ell_2}^2 = O_P \left(\frac{\bar{d}_n s' \log m_n}{n} \right) \quad (4.40)$$

Therefore, when $s' \leq s_n$, the ℓ_1 - ℓ_q regression estimator achieves the optimal rate of convergence.

Remark 4.9. Generally, the conditions for estimation consistency is weaker than those for variable selection consistency. For $q = 1$, why assumption 2 and 1 are weaker than the assumptions in theorem 3.1 can be found in (Meinshausen and Yu, 2006) and (Bickel et al., 2007). The cases for $q > 1$ and the group cases should follow in a similar way.

V. RISK CONSISTENCY

In this section, we study the risk consistency (or persistency) property with random design, which holds under a much weaker condition than variable selection consistency and does not need the true model to be linear. Instead of directly to show the persistency result for the estimator defined in equation 2.1, we show the persistency result for a constrained form estimator, which is equivalent to the estimator in 2.1 in the sense of primal and dual problems.

Due to the fact of random design and increasing dimensions, the same triangular array statistical paradigm as in (Greenshtein and Ritov, 2004) is adopted. In the following, we use calligraphic letter, such as \mathcal{Z} to represent random variables, while Z to represent its realization. Consider the triangular array $\mathcal{Z}_1^{(n)}, \dots, \mathcal{Z}_n^{(n)}$ (which is simplified as $\mathcal{Z}_1, \dots, \mathcal{Z}_n$),

our study mainly focus on the case where $\mathcal{Z}_1, \dots, \mathcal{Z}_n \stackrel{iid}{\sim} F_n \in \mathcal{F}^n$, where \mathcal{F}^n is a collection of distributions of $m_n + 1$ dimensional i.i.d. random vectors

$$\mathcal{Z}_i = (\mathcal{Y}_i, \mathcal{X}_{i,\underline{1}}, \dots, \mathcal{X}_{i,\underline{m_n}}) \quad i = 1, \dots, n \quad (5.1)$$

with the corresponding realizations

$$Z_i = (Y_i, X_{i,\underline{1}}, \dots, X_{i,\underline{m_n}}) \quad i = 1, \dots, n. \quad (5.2)$$

Denote

$$\gamma = (-1, \beta_{\underline{1}}, \dots, \beta_{\underline{m_n}}) = (\beta_{\underline{0}}, \beta_{\underline{1}}, \dots, \beta_{\underline{m_n}}), \quad (5.3)$$

and define

$$R_{F_n}(\beta) = \mathbb{E} \left(\mathcal{Y} - \sum_{\underline{j}=1}^{m_n} \mathcal{X}_{\underline{j}} \beta_{\underline{j}} \right)^2 = \gamma^T \Sigma_{F_n} \gamma \quad (5.4)$$

where $\mathcal{Z} = (\mathcal{Y}, \mathcal{X}_{\underline{1}}, \dots, \mathcal{X}_{\underline{m_n}}) \sim F_n \in \mathcal{F}^n$ and $(\Sigma_{F_n}) = \mathbb{E} \mathcal{Z}^T \mathcal{Z}$.

Given n observations Z_1, \dots, Z_n , denote their empirical distribution by \hat{F}_n and define the empirical risk as

$$R_{\hat{F}_n}(\beta) = \gamma^T \Sigma_{\hat{F}_n} \gamma \quad (5.5)$$

where $\Sigma_{\hat{F}_n} = \frac{1}{n} \sum_{i=1}^n Z_i Z_i^T$.

Given a sequence of sets of predictors $\mathcal{B}_n = \{\sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q} \leq L_n\}$, the sequence of estimators $\hat{\beta}^{\hat{F}_n}$ is called persistent if for every sequence $F_n \in \mathcal{F}^n$,

$$R_{F_n}(\hat{\beta}^{\hat{F}_n}) - R_{F_n}(\beta_*^{F_n}) \xrightarrow{P} 0, \quad (5.6)$$

where

$$\hat{\beta}^{\hat{F}_n} = \arg \min_{\beta \in \mathcal{B}_n} R_{\hat{F}_n}(\beta) = \arg \min_{\beta \in \mathcal{B}_n} \|Y - X\beta\|_{\ell_2}^2 \quad (5.7)$$

$$\beta_*^{F_n} = \arg \min_{\beta \in \mathcal{B}_n} R_{F_n}(\beta). \quad (5.8)$$

To show the persistency result, a moment condition as in (Zhou et al., 2007) is needed.

Assumption 3 For each $\underline{j}, \underline{k} \in \{1, \dots, m_n + 1\}$, denote $E = (\mathcal{Z} \mathcal{Z}^T - \mathbb{E}(\mathcal{Z} \mathcal{Z}^T))_{\underline{j}, \underline{k}}$, where $\mathcal{Z} = (\mathcal{Y}, \mathcal{X}_{\underline{1}}, \dots, \mathcal{X}_{\underline{m_n}})$, suppose that there exists some constants M and s .

$$\mathbb{E}(|E|^q) \leq q! M^{q-2} s/2 \quad (5.9)$$

for every $q \geq 2$ and every $F_n \in \mathcal{F}^n$.

Theorem 5.1. Suppose that $m_n \leq e^{n^\xi}$ for some $\xi < 1$. If $L_n = o((n/\log n)^{1/4})$, then ℓ_1 - ℓ_q regularized regression is persistent. That is, for every sequence $F_n \in \mathcal{F}^n$:

$$R_{F_n}(\hat{\beta}^{F_n}) - R_{F_n}(\beta_*^{F_n}) = o_P(1). \quad (5.10)$$

Proof: For any $\underline{j}, \underline{k} \in \{1, \dots, m_n + 1\}$ and any $\delta > 0$, from assumption 3 we can apply the Bernstein's inequality and obtain

$$\mathbb{P}\left(|(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| > \delta\right) \leq e^{-cn\delta^2} \quad (5.11)$$

for some $c > 0$. Therefore, by Bonferoni bound we have

$$\mathbb{P}\left(\max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| > \delta\right) \leq m_n^2 e^{-cn\delta^2} \leq e^{2n^\xi - cn\delta^2} \leq e^{-cn\delta^2/2} \quad (5.12)$$

for large enough n . For a sequence $\delta_n = \sqrt{\frac{2 \log n}{cn}}$, we have

$$\mathbb{P}\left(\max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| > \delta_n\right) \leq \frac{1}{n} \rightarrow 0 \quad (5.13)$$

which implies that

$$\max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| = O_P\left(\sqrt{\frac{\log n}{n}}\right). \quad (5.14)$$

Therefore,

$$\sup_{\beta \in \mathcal{B}_n} |R_{F_n}(\beta) - R_{\hat{F}_n}(\beta)| = \sup_{\beta \in \mathcal{B}_n} |\gamma^T (\Sigma_{F_n} - \Sigma_{\hat{F}_n}) \gamma| \quad (5.15)$$

$$\leq \max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| \|\gamma\|_{\ell_1}^2 \quad (5.16)$$

$$\leq \max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| \left(1 + \sum_{j=1}^{p_n} \|\beta_j\|_{\ell_1}\right)^2 \quad (5.17)$$

$$\leq \max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| \left(1 + \sum_{j=1}^{p_n} (d_j)^{1/q'} \|\beta_j\|_{\ell_q}\right)^2 \quad (5.18)$$

$$\leq \max_{\underline{j}, \underline{k}} |(\Sigma_{\hat{F}_n})_{\underline{j}, \underline{k}} - (\Sigma_{F_n})_{\underline{j}, \underline{k}}| (1 + L_n)^2 = o_P(1)$$

for $L_n = o((n/\log n)^{1/4})$.

Further, by definition, we have $R_{\hat{F}_n}(\hat{\beta}^{F_n}) \leq R_{F_n}(\beta_*^{F_n})$, combining with the following inequalities

$$R_{F_n}(\hat{\beta}^{F_n}) - R_{\hat{F}_n}(\hat{\beta}^{F_n}) \leq \sup_{\beta \in \mathcal{B}_n} |R_{F_n}(\beta) - R_{\hat{F}_n}(\beta)| \quad (5.19)$$

$$R_{\hat{F}_n}(\beta_*^{F_n}) - R_{F_n}(\beta_*^{F_n}) \leq \sup_{\beta \in \mathcal{B}_n} |R_{F_n}(\beta) - R_{\hat{F}_n}(\beta)|. \quad (5.20)$$

This implies that

$$R_{F_n}(\widehat{\beta}^{\widehat{F}_n}) - R_{F_n}(\beta_*^{F_n}) \leq 2 \sup_{\beta \in \mathcal{B}_n} |R_{F_n}(\beta) - R_{\widehat{F}_n}(\beta)| = o_P(1), \quad (5.21)$$

which completes the proof. \square

VI. DISCUSSIONS

The results presented here show that many good properties from ℓ_1 -regularization (Lasso) naturally carry on to the ℓ_1 - ℓ_q cases ($1 \leq q \leq \infty$), even if the number of variables within each group also increase with the sample size n . Using fixed design, we get both variable selection and estimation consistency under different conditions. Using random design, we get persistency under a much weaker condition. Our results provide a unified treatment for both the iCAP estimator ($q = \infty$) and the group Lasso estimator ($q = 2$).

Our results can also provide theoretical analysis to the simultaneous Lasso estimator (Turlach et al., 2005; Tropp et al., 2006) for joint sparsity. Which can find a good approximation of several response variables at once using different linear combinations of the high dimensional covariates. At the same time, it tries to balance the error in approximation against the total number of covariates that participate. Assuming that we have altogether \bar{d}_n response, the i -th signal is represented as $Y^{(i)} \in \mathbb{R}^n$, and the design matrix is $X = (X_{\underline{1}}, \dots, X_{\underline{p_n}}) \in \mathbb{R}^{n \times p_n}$. Denote the model as

$$Y^{(i)} = X\beta^{(i)} + \epsilon^{(i)}, \quad i = 1, \dots, \bar{d}_n \quad (6.1)$$

The simultaneous Lasso estimator can be formulated as

$$\widehat{\beta}^{(1)}, \dots, \widehat{\beta}^{(\bar{d}_n)} = \arg \min_{\beta^{(1)}, \dots, \beta^{(\bar{d}_n)}} \frac{1}{2n} \sum_{k=1}^{\bar{d}_n} \|Y^{(k)} - X\beta^{(k)}\|_{\ell_2}^2 + \lambda_n \sum_{j=1}^{p_n} \max_{\underline{\ell} \in \{1, \dots, \bar{d}_n\}} |\beta_{\underline{\ell}}^{(j)}|, \quad (6.2)$$

This problem can be formulated as a standard ℓ_1 - ℓ_q regularized regression estimator with $q = \infty$. For this, define

$$\widetilde{Y} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(\bar{d}_n)} \end{pmatrix} \in \mathbb{R}^{n\bar{d}_n} \quad \widetilde{X} = I_{\bar{d}_n} \otimes X = \begin{pmatrix} X & & \\ & \ddots & \\ & & X \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta^{(1)} \\ \vdots \\ \beta^{(\bar{d}_n)} \end{pmatrix} \quad (6.3)$$

where \otimes denotes the Kronecker product. Therefore, the simultaneous Lasso estimator can be rewritten as

$$\widehat{\beta}^{(1)}, \dots, \widehat{\beta}^{(\bar{d}_n)} = \arg \min_{\beta^{(1)}, \dots, \beta^{(\bar{d}_n)}} \frac{1}{2n} \|\widetilde{Y} - \widetilde{X}\beta\|_{\ell_2}^2 + \lambda'_n \sum_{j=1}^{p_n} (\bar{d}_n) \max_{\underline{\ell} \in \{1, \dots, \bar{d}_n\}} |\beta_{\underline{\ell}}^{(j)}| \quad (6.4)$$

where $\lambda'_n = \lambda_n/\bar{d}_n$. This is just an ℓ_1 - ℓ_∞ regularized regression estimator with block design. Therefore, all results in this paper can be applied to analyze such type estimators.

VII. ACKNOWLEDGEMENTS

We thank John Lafferty, Pradeep Ravikumar, Alessandro Rinaldo, Larry Wasserman, and Shuheng Zhou for their very helpful discussions and comments.

REFERENCES

- AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* 267–281.
- BACH, F. (2007). Consistency of the group lasso and multiple kernel learning. Tech. rep. ArXiv:0707.3390.
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. (2007). Simultaneous analysis of lasso and dantzig selector. *Technical report, U.C.Berkeley* .
- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Press.
- BUNEA, F., ALEXANDRE, B., TSYBAKOV, A. and WEGKAMP, M. (2007a). Aggregation for gaussian regression. *The Annals of Statistics* **35** 1674–1697.
- BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007b). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169–194.
- CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific and Statistical Computing* **20** 33–61.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics* **32** 407–499.
- FU, W. and KNIGHT, K. (2000). Asymptotics for lasso type estimators. *The Annals of Statistics* **28** 1356–1378.
- GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Journal of Bernoulli* **10** 971–988.
- LEDoux, M. and TALAGRAND, M. (1991). *Probability in Banach spaces: isoperimetry and processes*. Springer-Verlag Inc.
- MALLOWS, C. L. (1973). Some comments on C_p . *Technometrics* **15** 661–675.

- MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2007). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B, Methodological* **70** 53–71.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.
- MEINSHAUSEN, N. and YU, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. Tech. Rep. 720, Department of Statistics, UC Berkeley.
- OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9** 319–337.
- RAVIKUMAR, P., LIU, H., LAFFERTY, J. and WASSERMAN, L. (2007). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems 20*. MIT Press.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological* **58** 267–288.
- TROPP, J., GILBERT, A. C. and STRAUSS, M. J. (2006). Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing* **86** 572–588.
- TURLACH, B., VENABLES, W. N. and WRIGHT, S. J. (2005). Simultaneous variable selection. *Technometrics* **27** 349–363.
- WAINWRIGHT, M., RAVIKUMAR, P. and LAFFERTY, J. (2006). High-dimensional graphical model selection using ℓ_1 -regularized logistic regression. In *Advances in Neural Information Processing Systems 19*. MIT Press.
- WAINWRIGHT, M. J. (2006). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programs. In *Proc. Allerton Conference on Communication, Control and Computing*.
- YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B, Methodological* **68** 49–67.
- ZHAO, P., ROCHA, G. and YU, B. (2008). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics* **to appear**.
- ZHAO, P. and YU, B. (2007). On model selection consistency of lasso. *J. of Mach. Learn. Res.* **7** 2541–2567.
- ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2007). Compressed regression. Tech. rep., Carnegie Mellon. Technical report.
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.